

<https://helda.helsinki.fi>

The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): plastid genome structure evolution in Solanaceae

Amiryousefi, Ali

2018-04-25

Amiryousefi , A , Hyvönen , J T & Poczai , P 2018 , ' The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): plastid genome structure evolution in Solanaceae ' , PLoS One , vol. 13 , no. 4 , 0196069 . <https://doi.org/10.1371/journal.pone.0196069>

<http://hdl.handle.net/10138/235198>

<https://doi.org/10.1371/journal.pone.0196069>

cc_by

publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

RESEARCH ARTICLE

The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae

Ali Amiryousefi¹, Jaakko Hyvönen^{1,2}, Péter Poczai^{2*}

1 Organismal Evolutionary Biology Research Program, Faculty of Biology and Environmental Sciences, Viikki Plant Science Centre, University of Helsinki, Helsinki, Finland, **2** Finnish Museum of Natural History (Botany), University of Helsinki, Helsinki, Finland

* peter.poczai@gmail.com



Abstract

Bittersweet (*Solanum dulcamara*) is a native Old World member of the nightshade family. This European diploid species can be found from marshlands to high mountainous regions and it is a common weed that serves as an alternative host and source of resistance genes against plant pathogens such as late blight (*Phytophthora infestans*). We sequenced the complete chloroplast genome of bittersweet, which is 155,580 bp in length and it is characterized by a typical quadripartite structure composed of a large (85,901 bp) and small (18,449 bp) single-copy region interspersed by two identical inverted repeats (25,615 bp). It consists of 112 unique genes from which 81 are protein-coding, 27 tRNA and four rRNA genes. All bittersweet plastid genes including non-functional ones and even intergenic spacer regions are transcribed in primary plastid transcripts covering 95.22% of the genome. These are later substantially edited in a post-transcriptional phase to activate gene functions. By comparing the bittersweet plastid genome with all available Solanaceae sequences we found that gene content and synteny are highly conserved across the family. During genome comparison we have identified several annotation errors, which we have corrected in a manual curation process then we have identified the major plastid genome structural changes in Solanaceae. Interpreted in a phylogenetic context they seem to provide additional support for larger clades. The plastid genome sequence of bittersweet could help to benchmark Solanaceae plastid genome annotations and could be used as a reference for further studies. Such reliable annotations are important for gene diversity calculations, synteny map constructions and assigning partitions for phylogenetic analysis with de novo sequenced plastomes of Solanaceae.

OPEN ACCESS

Citation: Amiryousefi A, Hyvönen J, Poczai P (2018) The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. PLoS ONE 13 (4): e0196069. <https://doi.org/10.1371/journal.pone.0196069>

Editor: Berthold Heinze, Austrian Federal Research Centre for Forests BFW, AUSTRIA

Received: April 24, 2017

Accepted: April 5, 2018

Published: April 25, 2018

Copyright: © 2018 Amiryousefi et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The chloroplast genome data is available from NCBI under accession number KY863443.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The genus *Solanum* L., with approximately 1,400 species, is one of the largest genera of angiosperms, and includes many major and minor food crops such as tomato, potato, eggplant, and pepino. Bittersweet (*Solanum dulcamara* L.) is a European native diploid ($2n = 2 \times = 24$)



Fig 1. The berries and flowers of *Solanum dulcamara* L.

<https://doi.org/10.1371/journal.pone.0196069.g001>

species, which is found throughout the northern hemisphere across a wide range of habitats. It was also introduced to North America possibly for its medicinal properties [1]. It is still used as a source of various alkaloids with diuretic, diaphoretic properties to treat rheumatism and skin diseases in Asia and India [2, 3].

This semi-woody perennial vine is easy to recognize (Fig 1). However, it is a highly polymorphic and phenotypically plastic species showing extreme forms, which has led to confused taxonomy. Previous treatments placed *Solanum dulcamara* to sect. *Dulcamara* (Moench) Dumort. in subg. *Potatoe* (G.Don) D'Arcy related to potatoes (sect. *Petota* Dumort.) and tomatoes (sect. *Lycopersicum* (Tourn.) Wettst.) [4–7]. This was based on scandent habit, pinnate leaves and on the articulation of pedicels above the base [1, 4]. However, recent phylogenetic studies have shown that it belongs to the Dulcamaroid clade [8–11], which is closely related to the Morelloid clade including species of black nightshades of sect. *Solanum* (e.g. *S. nigrum* L. and *S. scabrum* Mill.).

Solanum dulcamara serves as a host for important plant pathogens such as those causing bacterial wilt (*Ralstonia solanacearum* (Smith 1896) Yabuuchi et al. 1996), late blight (*Phytophthora infestans* (Mont.) de Bary.) and also for some viruses [12, 13]. Late blight, is one of the most serious potato diseases worldwide [14]. However, it was shown that bittersweet has a minimal role in late blight infections since most plants are resistant and the inocula of the pathogen do not

overwinter [15]. Populations of this species seem to have experienced a genetic bottleneck [16], but some allelic variation was found to be distributed among populations resulting in more structured populations at larger regional levels [17]. The differentiation of the populations could have arisen by genetic drift or even by inbreeding over a very long period. Bittersweet is mostly an outcrossing species, but its population structure might have been affected by its perennial self-compatibility [18], reducing genetic diversity within regional populations and enhancing inbreeding. This leads to high interpopulation or spatial differentiation [17]. Genetic drift, on the other hand, may not have shaped the population structure of the species recently based on the observed moderate level of diversity among populations [16, 17]. However, over a longer time scale population expansion from postglacial refugia is known to leave such traces [19].

High throughput sequencing is revolutionizing phylogenetics as it allows to obtain hundreds to thousands of markers in a cost effective way. Complete plastid genome (plastome) sequences now could be easily acquired for phylogenomic analyses with relatively low cost. Angiosperm plastid genomes exist in circular and linear forms [20] and the percentage of each form varies within plant cells [21]. They are small, typically ~ 120–150 kb in size and have a highly conserved quadripartite structure containing two inverted repeats (IRA and IRB), which separate the large and small single copy regions (LSC and SSC). The plastid genome includes 110–130 genes primarily participating in photosynthesis, transcription and translation [22]. Their conserved gene content, order and organization makes them relatively well suited for evolutionary studies since gene losses, structural rearrangements, pseudogenes or additional mutation events could be characteristic for some lineages. The information from length mutational events could be used in addition to the information from DNA substitutions occurring in the plastid genome. Such changes have been shown to be informative for example in Araliaceae [23], Geraniaceae [24], Poaceae [25] and in early embryophyte lineages [26]. It has been shown that independent gene and intron losses are limited to the more derived monocot and eudicot clades with lineage-specific correlation between rates of nucleotide substitutions, indels, and genomic rearrangements [27].

Here we present the complete chloroplast genome sequence of bittersweet using high-throughput sequencing, as well as the assembly, annotation, gene expression and unique structure characterization of its plastome. We also compare the gene order, inverted repeat (IR) length and examine the variation of structural changes across the family. In order to achieve this we revise the annotations of Solanaceae plastid genome records and correct possible errors. Using this edited plastid genome dataset we present a phylogenetic hypothesis of Solanaceae and examine the distribution of structural changes in the plastid genomes.

Materials and methods

Chloroplast isolation

Bittersweet leaves were collected in the Kaisaniemi Botanical Garden of the University of Helsinki, Finland during the summer of 2015. DNA isolation was carried out according to the modified high-salt protocol of Shi et al [28]. DNA concentration was measured with a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and checked on 0.8% agarose gel. We carried out a multiply-primed rolling circle amplification (RCA) according to the protocol of Atherton et al. [29] using a REPLI-g Mini Kit (Qiagen, Hilden, Germany) to produce abundant DNA template.

Plastid genome sequencing

Paired-end libraries of 300 bp were prepared with Illumina TruSeq DNA Sample prep kit (Illumina, San Diego, CA, USA). Fragment analysis was conducted with an Agilent Technologies

2100 Bioanalyzer using a DNA 1000 chip. Sequencing was carried out on an Illumina MiSeq platform from both ends with 150 bp read length.

Genome assembly and annotation

Raw reads were first filtered to obtain high-quality clean data by removing low quality reads with a sliding window quality cutoff of Q20 using Trimmomatic [30]. Plastid reads were filtered by reference mapping to Solanaceae plastid genome sequences using Geneious 9.1.7. [31] with medium-low sensitivity and 1,000 iterations. From the collected reads a *de novo* assembly was carried out with the built-in Geneious assembler platform with zero mismatches and gaps allowed among the reads. The similar procedure was conducted with Velvet v1.2.10 [32] with k-mer length 37, minimum contig length 74 and default settings by applying a 400× upper coverage limit. The resulting contigs were then circularized by matching end points. The results of the reference mapping and two *de novo* methods were compared and inspected. Sanger-based gap closure and IR junction verification was carried out following Moore et al. [33]. Gene annotation was made with a two-step procedure. First we used gene prediction tools DOGMA [34], tRNAscan-SE [35], cpGAVAS [36], Verdant [37] and GeSeq [38] to obtain annotations based on different approaches. In a second step we inspected and curated all annotation manually with comparisons to all published (as of 18.10.2016) plastid genomes of Solanaceae using Geneious. Local BLAST searches were further carried out to confirm the position of CDS regions and genes. We confirmed start and stop codons manually and by comparison to RNA-seq data. For each gene we inspected gene length based on amino acid translations and reconfirmed any internal stop codons. The resulting genome map was drawn with OGDRAW v.1.2 [39]. The annotated bittersweet plastid genome was further used as a reference to revise all Solanaceae plastid genomes (deposited by 16.8.2016). Reannotation followed the two-step protocol described above. Plastid genome sequences were transformed into fasta file format then annotated with the software tools [34–38]. All annotations were transferred to Geneious as a new track under the corresponding genome. Sequences were aligned, compared and manually curated compared to bittersweet.

Genome analyses

Codon frequency and relative synonymous codon usage (RSCU) was calculated on the basis of protein-coding genes using an *in-house* script. We also computed the overall mean of pairwise distances of 80 protein-coding genes of the 32 Solanaceae species based on the Kimura 2-parameter model using MEGA 7.0.21 [40]. Standard error estimate(s) were obtained using bootstrap (1,000 replicates). Complete plastid genome sequences were compared and aligned using mVISTA online tools [41], while the expansion and contraction of the inverted repeat (IR) regions at junction sites was examined and plotted using IRscope [42]. We identified and located repeat sequences ($n \geq 30$ bp and a sequence identity $\geq 90\%$) found in the bittersweet plastome using REPuter [43]. Repeats larger than 10 bp were classified into the following groups: (i) forward or direct repeats (F), (ii) repeats found in reverse orientation (R), (iii) palindromic repeats forming hairpin loops in their structure (P) and (iv) repeats found in reverse complement orientation (C). Because REPuter overestimates the number of repeats we manually inspected the output file and located the repeats in Geneious. Redundant repeats found entirely within other repeats as well as duplicated parts of tRNAs were pruned. Perfect and compound simple sequence repeats (SSRs) interrupted by 100-bp were located with MISA [44]. A threshold level of seven was applied to mononucleotide repeats, four to dinucleotide repeats and three to tri-, tetra-, penta-, and hexanucleotide repeats. Output files were manually edited and exported to Geneious for further inspection.

Transcriptome analysis and RNA editing site prediction

RNA-seq library files were downloaded from NCBI Short Read Archive for *Solanum dulcamara* (SRR2056039). Reads were mapped to the complete plastid genome and filtered reads were collected with Bowtie 2.0 [45] (mismatch ≤ 2). RNA-seq reads were re-mapped with Geneious using the genome annotation to calculate reads per kilobase per million (RPKM), fragments per kilobase of exon per million fragments mapped (FPKM) and transcripts per million (TPM) for transcript variants. Ambiguously mapped reads were counted as partial matches for each CDS. Putative RNA Editing sites were predicted with an *in silico* approach using the PREP database [46]. Verification of the predicted editing sites was carried out by FreeBayes [47] variant calling.

Phylogenomic analyses

Our aim was to compare the 32 chloroplast genomes of Solanaceae (data present in NCBI on 16.8.2016) with each other and try to hypothesize when changes have taken place between/among the species and major clades. As outgroup terminals we used *Coffea arabica* L. of Rubiaceae, *Ipomoea batatas* (L.) Lam. and *I. purpurea* (L.) Roth. We aligned the 35 complete chloroplast genomes (S1 Table) with MAFFT [48] (S1 Data) since they were lacking inversions or other major changes. We conducted maximum likelihood (ML) analyses using RAXML-NG [49] under three different strategies. 1) One of the IR regions was removed from all plastid genomes to reduce overrepresentation of duplicated sequences then we run RAXML-NG on the unpartitioned alignment under GTR+I+G substitution model as a single partition; 2) The same data matrix was partitioned by gene, exon, intron and intergenic spacer regions ($n = 258$) and allowed separate base frequencies, α -shape parameters, and evolutionary rates to be estimated for each; 3) we inferred the best-fitting partitioning strategy with PartitionFinder2 [50] for the alignment ($n = 24$). The best fitting nucleotide substitution models were inferred with jModelTest2 [51]. Branch support values were obtained from 10,000 non-parametric bootstrapping. For each alignment we conducted ten separate runs with RAXML-NG v0.5.0b since log-likelihoods could show variation among individual runs [52]. The complete plastid genome alignment was analyzed also with parsimony as an optimality criterion using the program TNT [53]. The matrix included 19,956 parsimony informative characters and due to its small size we were able to perform analyses using “traditional” search starting from Wagner trees improved using tree bisection reconnection (TBR) algorithm. This search was performed twice with 3,000 replications. We also examined the phylogenetic distribution of structural changes using the tree constructed with parsimony and ML methods implemented in the ancestral state reconstruction tools of Mesquite 3.2 [54]. Major genomic changes were binary coded (S2 Data) and mapped on phylogenetic trees. Phylogenetic trees were visualized and edited with TreeGraph2 [55].

Results and discussion

Chloroplast genome assembly and validation

Enriched chloroplast DNA was used to generate 1,645,956 paired-end reads, with an average fragment length of 277 bp, which generated average $1,340 \times$ genome coverage. Low quality reads (Q20) were filtered out, and the remaining high quality reads were utilized in further assembly. For genome assembly we used one reference mapping and two *de novo* methods. As a first step quality filtered reads were mapped to Solanaceae reference genomes, which resulted in an entire contig showing good agreement with published genome sequences. Based on these collected reads we used Geneious and Velvet to produce a single contiguous fragment

representing the plastid genome. The three assemblies were compared and discrepancies were manually resolved. With Velvet we obtained a linear contig 43 bp longer (155,623 bp) than with Geneious (155,580 bp) which was caused by a repeated sequence at the start and end point and these were removed. Most *de novo* methods do not account for the circularity of the plastid genome, while Geneious overcomes this by allowing contig circularization during the assembly. The assembly was validated by PCR amplification and Sanger sequencing targeting the four junctions between the IRs and LSC/SSC regions. Sanger results showed identical sequences when compared to the plastid genome demonstrating the accuracy of the assembly. The final chloroplast genome sequence was then submitted to GenBank (KY863443).

Genome organization, repeats and sequence diversity

The chloroplast genome of *Solanum dulcamara* is 155,580 bp long showing a quadripartite structure of long and small single-copy regions of length 85,901 and 18,449 bp, separated with two inverted repeat regions of 25,615 bp (Fig 2). The genome contains 81 protein-coding, 27 tRNA and four rRNA genes comprising the total of 114 unique genes (S2 Table). Seventeen genes contained introns, with *ycf3* and *clpP* containing two. All of these belong to group II introns except *trnL-UAA* with group I intron (S3 Table). The distribution of the genes on different regions of the genome exhibit similarity with other Solanaceae with 13 genes in the SSC and 19 genes in the IR while the rest were on the LSC. The overall GC content of the chloroplast genome is 37.8% resembling other species of Solanaceae (S4 Table). Eighty percent of the total length of the genome is related to genetic regions. The Arg amino acid coded with AGA codon was the most frequent codon showing RSCU rate of 1,187 (S5 Table).

The majority of the genes show relatively slow evolutionary divergence since all genes had an average sequence distance of less than 0.10 (S6 Table). Low levels of sequence distances indicate the conserved nature of protein-coding genes in Solanaceae. The only gene showing slightly larger distance with a unique function was *sprA* ($d = 0.114$; $S.E = 0.016$). Chloroplast genes are mostly subjected to purifying selection and low sequence diversity is due to conservation of the functions of the photosynthetic system. In this context the plastid genome diversity of Solanaceae do not resemble other economically important plant families such as Poaceae where plastid genomes harbor many divergent genes and unique plastid rearrangements [25].

Using MISA we identified 374 SSRs in the bittersweet plastid genome, of which 253 were mono-, 40 di-, 70 tri-, 10 tetra- and one was a pentanucleotide (S7 Table and S3 Data). SSRs were more abundant in the LSC and SSC regions compared to the IRs and 107 occurred in compound formation that were composed of several combinations of SSRs interrupted by maximum distances of 100 bp. The most abundant motifs of the SSRs were poly-A/T stretches characteristic of angiosperm plastid genomes. We also identified 25 larger repeats (> 10 bp) in the bittersweet plastid genome composed of 12 forward, five reverse, five palindromic and three mixed (forward/palindromic) repeats (Table 1) using REPuter. The largest repeat with a size of 83 bp was a forward repeat found in the IGS region of *ycf3* and *trnS-GGA*. Forward repeats were commonly distributed in the intergenic spacer regions of the genome located mostly in the LSC. Two repeats were found among the introns of *ndhA*, *ycf3* and *petD* while one repeat appeared in the *infA* pseudogene. Three repeats were found among the CDS of *atpI*, *ndhC* and *ycf2*, while another motif was repeated in the *psaA* and *psbB* gene. The repeats in *atpI* and *ycf2* seem to be conserved since they have also been reported from grasses [25]. The most variable region was the *trnE-UUC*—*trnT-GGU* IGS, which had two palindromic and one forward repeat.



Fig 2. Map of the chloroplast genome of the *Solanum dulcamara*. Genes lying inside of the outer circle are transcribed counterclockwise while those outside that circle are transcribed clockwise. Genes belonging to different functional groups are color coded differently and the GC, AT content of the genome are plotted on the inner circle as dark and light gray, respectively. The inverted repeats, large single copy, and small single copy regions are denoted by IR, LSC, and SSC, respectively.

<https://doi.org/10.1371/journal.pone.0196069.g002>

Reannotation of Solanaceae plastid genomes

We noticed a litany of errors in currently deposited annotations, which were corrected for our analyses in a two-step curation process using gene prediction tools followed by manual adjustments. The reannotated genome files could be accessed as an online supplement (S4 and S5 Data). We provide here the first annotation for the sequences of *S. pennellii* Correll and

Table 1. Repeat sequences of the *Solanum dulcamara* chloroplast genome.

No	Type	Location		Region	Repeat unit	Period size (bp)	Copy Nr.
1	F	<i>ycf3</i> — <i>trnS</i> -GGA	IGS	LSC	AACAATTTTAAAGAAAAATGTATCTTTATCCCGGAGTC TTGAAGGAAAGAAAAATGGTTCTTTGTTTGGACTTTGATGAAA	83	2
2	F	<i>psaA</i> and <i>psbB</i>	CDS	LSC	TGCAATAGCTAAATGGTGATGGGCAATATCAGTCAGCC	38	2
3	F	<i>ndhA</i> and <i>ycf3</i>	intron	LSC/ SSC	CAGAACCGTACGTGAGATTTTCACCTCATACGGCTCCT	38	2
4	F	<i>infA</i>	pseudogene	LSC	AGGTATCAACTAATCTAATCCAATTTGGATATTATAAA	38	2
5	F	<i>atpB</i> — <i>rbcL</i>	IGS	LSC	TTAGCACTCGATGAGACTGAGTTAATTGCAAGCT	34	2
6	F	<i>psbA</i> — <i>ycf3</i> — <i>trnS</i> -GGA	IGS	LSC	TTAATATAATAAAAAGAAGTCTATTTTGT	29	2
7	F	<i>sprA</i> — <i>trnL</i> -UAG	IGS	SSC	CCTTTTAACTCTATTCTTAATTGAGT	28	2
8	P	<i>rps12</i> — <i>trnV</i> -GAC	IGS	IR	TGAGATTTTCACCTCATACGGCTCCT	26	2
9	P	<i>petD</i>	intron	LSC	TATAAGTGAAGTAGATAAAACGGAAT	26	2
10	F	<i>trnG</i> -GCC— <i>trnR</i> -UCU	IGS	LSC	TTAGTACATCATTGAATATACAA	23	2
11	F	<i>psaJ</i> — <i>rpl33</i>	IGS	LSC	GTGGACGGGCTGAGGAATGGGG	22	2
12	F/P	<i>rps12</i> — <i>trnV</i> -GAC	IGS	IR	ATTAGATTAGTATTAGTTAGT	21	4
13	F	<i>ndhC</i> — <i>trnV</i> -UAC	IGS	LSC	TCCTTTTATTATTATTTAAT	20	2
14	P	<i>psbT</i> — <i>psbN</i>	IGS	LSC	AGTTGAAGTACGGAGCCTCC	20	2
15	F	<i>trnE</i> -UUC— <i>trnT</i> -GGU and <i>rps4</i> — <i>trnT</i> -UGU	IGS	LSC	TTATTTAGTATTTCGAATT	19	2
16	F/P	<i>ycf2</i>	CDS	IR	CGATATTGATGATAGTGAC	19	4
17	F	<i>rps16</i> — <i>trnQ</i> -UUG	IGS	LSC	ATTATAATATTAATTA	16	3
18	P	<i>trnE</i> -UUC— <i>trnT</i> -GGU	IGS	LSC	TTTTATTAGAAA	13	2
19	P	<i>trnE</i> -UUC— <i>trnT</i> -GGU	IGS	LSC	CATCATACTATGA	13	2
20	R	<i>trnF</i> -GAA— <i>ndhJ</i>	IGS	LSC	TCTCCTCTTTT	11	2
21	R	<i>ndhC</i>	CDS	LSC	CATCAAAAACA	11	2
22	R	<i>atpH</i> — <i>atpI</i>	IGS	LSC	TTTATTATTTA	11	2
23	R	<i>atpI</i>	CDS	LSC	ACAAAAATAA	11	2
24	R	<i>petL</i> — <i>petG</i>	IGS	LSC	CCTCTTTTTT	10	2
25	F/P	<i>rps12</i> — <i>trnV</i> -GAC	IGS	IR	AACATAACT	10	6

<https://doi.org/10.1371/journal.pone.0196069.t001>

Lochroma loxense (Kunth) Miers, which entirely lacked genome features. A complete list of annotation errors is found in S8 Table, and illustrates the difficulties encountered when attempting to compare across genomes. These differences could cause considerable consequences inferring gene functionality or synteny. In general annotations of the LSC and SSC corresponding to the basic quadripartite structure of angiosperm plastid genomes were entirely missing or sparsely indicated. Inverted repeats (IRs) were either unannotated or their orientation, size and correct naming was erroneous. Compared to the tobacco reference order LSC-IRB-SSC-IRA [56], the erroneous annotation LSC-IRA-SSC-IRB is often applied. It is important to note that the IR sequences of the *Atropa belladonna* L. and *Saracha punctate* Ruiz. c Pav. were dissimilar. Inverted repeat sequences are under concerted evolution [22] and divergent sequences could be possible sequencing/assembly errors in these two genomes or they could represent a relatively rare case of chloroplast evolution. Several protein-coding genes had errors with assigned start/stop codons. For example, the start codon of the *rpoC2* gene is shifted with 12 bps in most deposited plastid genomes except in *Nicotiana* L. species and in *Datura stramonium* L. Annotations were found to be insufficient for genes containing introns since they were lacking exon and/or intron designations. The exon-intron boundaries had variable annotation for many genes with high level of synteny, e.g., *atpF* or *rpoC1*. Gene

annotations were missing for some species in case of *psbK* and *psbZ*, while the later was often annotated as *ihbA* now regarded as a synonym of *psbZ*.

Besides previously described genes we located and annotated hypothetical gene *ycf68* the 218 bp long small plastid RNA (*sprA*) gene in all studied genomes. Homologs of *sprA* are present in eudicots but absent from monocots and they are rarely annotated in plastid genomes. This gene was reported to play a role in the 16S rRNA maturation in *Nicotiana tabacum* L. [57], but its function is non-essential under normal growth conditions [58]. It is not part of the catalytic core nor does it guide the rRNA machinery rather it acts independently. In this respect its function is similar to other non-essential plastid sRNAs.

According to our experiences during the reannotation none of the currently existing tools provided submission ready annotations. They required minor or even extensive manual curation especially with the most commonly used DOGMA producing results which require expert interpretation and laborious adjustments. For example annotating intron-containing genes or genes with short exons such as *petB*, and dealing with trans-splicing reading frames like *rps12* is challenging with DOGMA. Moreover DOGMA [34] generates a special output file compared to CpGAVAS [36] or GeSeq [38], which generate standard general feature format (.gff) or GenBank (.gb) files that can be integrated with other software without further processing. From the currently available tools GeSeq [38] generated the highest quality results by annotating >95% of the genes and coding regions correctly compared to our curated reference set. In most cases annotation errors were propagated from erroneous references to newly assembled genomes creating a systematic problem in Solanaceae. For future reference we advise the jettison of outdated annotation tools such as DOGMA and advise the use of up-to-date novel software such as GeSeq to avoid complications. For de novo sequenced Solanaceae plastid genomes bittersweet can also serve as a novel reference for comparison and annotation.

Expansion and contraction of IR regions

By using the curated genome annotations we compared the junction sites of ten selected Solanaceae plastid genomes. In general IRs are systematically un-annotated in deposited plastid genomes with several genes, for example *rpl2*, missing. Pseudogenes like the truncated *ψrps19* are mislabeled or entirely missing, which made the comparison of the IR regions cumbersome and time consuming. Therefore, we utilized an *in house* script, IRscope [42] to overcome these problems, and located the IRs and plotted the genes in vicinity of the junctions (Fig 3). The length of the IR regions were similar ranging from 25,343 bp to 25,906 bp showing some expansion. The endpoint of the Solanaceae JLA is characteristically located upstream of the *rps19* and downstream of the *trnH*-GUG. In Solanoideae, the IR expanded to partially include *rps19* creating a truncated *ψrps19* copy at JLA, thus this pseudogene is missing from *Nicotiana*. The extent of the IR expansion to *rps19* varies from 24 to 91 bp and the end point seems to be conserved not exceeding to the following intergenic spacer region. Furthermore, *infA*, *ycf15*, and a copy of *ycf1* located on the JSB were detected as pseudogenes. In contrast to *Solanum tuberosum* and *S. lycopersicum* where JSB is tangent to the end of the pseudo *ycf1* gene, the copy of this gene in *S. dulcamara* is showing an extra part extended further to the SSC (Fig 3).

Phylogenetic relationships in Solanaceae

Our phylogenetic analyses of the whole plastid genome alignment resulted in highly resolved trees (Fig 4), with almost all clades recovered having maximum branch support values (S1 Fig). We conducted phylogenetic analysis with three different partitioning strategies under maximum likelihood and analyzed the matrix also using parsimony. All our analyses resolved similar topologies which confirm results of previous phylogenetic analyses based on fewer

Inverted Repeats

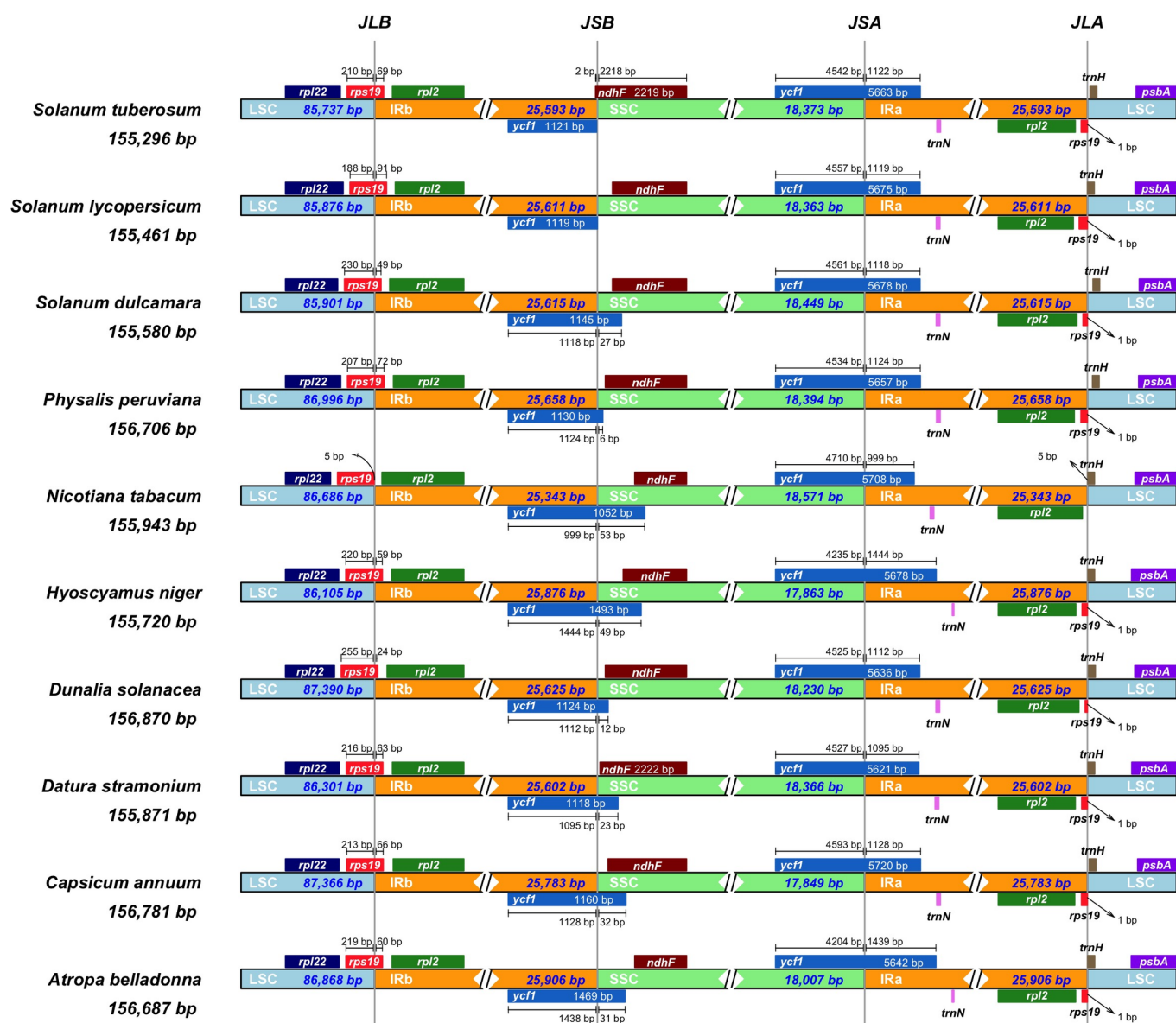


Fig 3. Junction sites of the inverted repeats. For each species, genes transcribed in positive strand are depicted on the top of their corresponding track with right to left direction, while the genes on the negative strand are depicted below from left to right. The arrows are showing the distance of the start or end coordinate of a given gene from the corresponding junction site. For the genes extending from a region to another, the T bar above or below them show the extent of their parts with their corresponding values in base pair while nothing is plotted for the genes tangent to the sites. The plotted genes and distances in the vicinity of the junction sites are the scaled projection of the genome. JLB (IRb /LSC), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (IRa/LSC) denote the junction sites between each corresponding two regions on the genome.

<https://doi.org/10.1371/journal.pone.0196069.g003>

genes [10, 59] but in several cases groups with low support values of earlier studies are resolved in our tree with high support values.

Trees of parsimony and ML analyses are congruent except for the clade composed of iochromas (S1 Fig). Iochrominae is a diverse clade of Physaleae with ca. 34 species and six

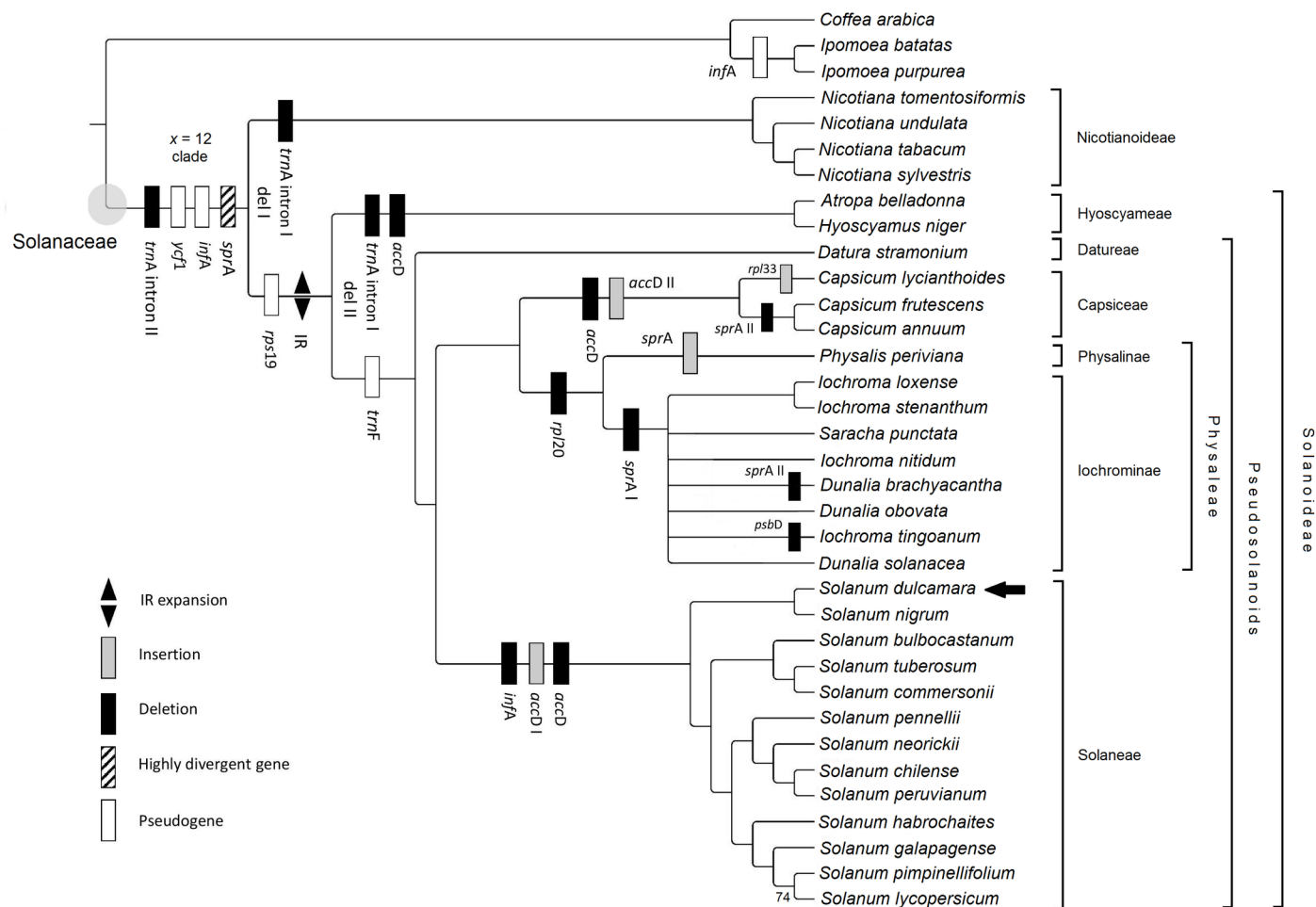


Fig 4. Cladogram illustrating the phylogenetic relationships of Solanaceae based on complete chloroplast genome sequences. Plastid genome rearrangement events are mapped on the branches of the best scoring maximum likelihood tree generated with RAXML-NG. Each node has 100% bootstrap support value. A node with lower support value indicated and those with support values below 50% collapsed. Currently recognized suprageneric groups are listed on the right.

<https://doi.org/10.1371/journal.pone.0196069.g004>

traditionally recognized genera, including *Acnistus* Schott, *Dunalia* Kunth, *Eriolarynx* (Hunz.) Hunz, *Iochroma* Benth., *Saracha* Ruiz & Pav. and *Vassobia* Rusby. Members of this group are shrubs of high elevation in the Andes displaying great diversity in floral characteristics and pollination system. Recent molecular phylogenetic studies resolved Iochrominae with high support value but relationships within the clade have remained poorly resolved [10, 59]. In this group nodal resolution does not scale proportionately to the length of sequence analyzed, and structural variations in the plastid genome seem to be accumulated as compared to other clades.

Iochrominae represented here by *Iochroma*, *Dunalia* and *Saracha* appear to be monophyletic based on the analyses of the complete chloroplast genome sequences. However, our results also suggest that two of these morphologically delimited genera (*Iochroma* and *Dunalia*) are not monophyletic. Smith and Baum [60] utilizing nuclear markers (ITS, *waxy* and LEAFY) also found that generic boundaries are not congruent with the current taxonomy. Iochromas might have highly reticulated history that is impossible to be represented by a dichotomic tree. The unequivocal resolution of iochromas will likely require the inclusion of nuclear genomic regions.

We resolved *Solanum dulcamara* in a separate clade with *S. nigrum* appearing as a sister group. This reinforces the close relationship of the Dulcamaroid and Morelloid clades as proposed by other molecular phylogenetic analyses based on fewer markers [8–10]. The informally named $x = 12$ clade is found in our analysis as sister to Nicotianoideae. In this group the chromosome numbers are based on 12 pairs [61], and members are estimated to have gone through two separate whole-genome duplication (WGD) events ca. 117 Ma [62] and 49 Ma BP [63], respectively. Increased sampling outside this group is needed since this could shed light on ancient WGDs in the family. Plastid genomes of Solanaceae hold much promise for resolving relationships among clades of the family that have previously been problematic. Although the phylogenomic tree presented in this study is largely robust it should be kept on mind that our sampling is still sparse in terms of the number of terminals. It is also important to note that organellar phylogenomics may fail in rapidly radiating groups with interspecific hybridization as exemplified here by *Ioichromas*. Other biological processes such as incomplete lineage sorting might also make phylogenetic analyses very difficult, however, organellar phylogenomics can be used to detect such processes.

Plastid genome structure of Solanaceae

Intending to identify and map the major structural changes of Solanaceae plastid genomes on the phylogenetic tree, we selected ten Solanaceae plastid genomes for detailed comparison representing diverse groups of the family and included two outgroup taxa in the analysis. Gene comparisons were extended to the entire Solanaceae dataset using local alignments with MAFFT and the curated genome annotations. The size of the plastid genomes varied between 155,312 bp (*Solanum tuberosum*) to 162,046 bp (*Ipomoea purpurea*) (S4 Table). Our comparison shows that gene content and synteny are highly conserved across Solanaceae plastid genomes (S2 Fig). All species analyzed display complete gene synteny when accounting for expansion and contraction of the IRs (Fig 3). The organization and evolution of Solanaceae plastid DNA have been analyzed by previous studies using restriction site methods [64], PCR surveys [65–68] and complete genome sequences [69–74]. These comparisons highlighted some features of Solanaceae but the phylogenetic distribution of these rearrangements have not been examined. Our comprehensive comparison of complete chloroplast genomes of ten Solanaceae and *S. dulcamara* confirm the presence of all the genomic rearrangements reported

Table 2. Major changes in the chloroplast genomes of Solanaceae.

Gene	Insertion	Deletion	Pseudogene	Notes
<i>accD</i>	2	1	-	24-bp deletion in the 'x = 12 clade' except (<i>Nicotiana</i> , <i>Datura</i> , <i>Physalis</i> , <i>Ioichromas</i>) 141-bp insertion in <i>Capsicum</i> 9-bp insertion in <i>Solanum</i>
<i>infA</i>	-	1	+	124-bp deletion in <i>Solanum</i>
<i>psbD</i>	-	1	-	64-bp deletion in <i>Ioichroma tongoanum</i>
<i>rpl20</i>	-	1	-	31-bp deletion Physaleae
<i>rpl33</i>	1		-	15-bp insertion in <i>Capsicum lycianthoides</i>
<i>rps19</i>	-	-	+	-
<i>sprA</i>	1	2	-	14-bp insertion in <i>Physalis</i> 52-bp deletion in <i>Capsicum</i> 37-bp deletion <i>Ioichromas</i>
<i>trnA-UGC</i>	-	2	-	108-bp and 141-bp intron deletion in <i>Nicotiana</i> and <i>Atropa/Hyoscyamus</i>
<i>trnF-GAA</i>	-	-	+	Uniting a group of Pseudosolanoids
<i>ycf1</i>	-	-	+	Truncated pseudogenization of one <i>ycf1</i> copy in Solanaceae.

<https://doi.org/10.1371/journal.pone.0196069.t002>

previously. We will briefly review the conclusions made before and then highlight the novel aspects resulting from our analysis and moreover, examine the distribution of these structural changes using the phylogenetic hypothesis constructed based on complete plastid genome alignment.

We observed ten characteristic features in Solanaceae plastid genomes linked to indels or pseudogenization processes (Table 2). Two genes, one copy of *ψycf1* and *ψrps19* at the IRb/SSC and IRa/LSC junction were truncated pseudogenes, while *infA* has become non-functional through partial degradation. The substitutions of *infA* orthologues in Solanaceae show almost equal numbers of substitutions at all codon positions with missing start codons. It is also a pseudogene in *Ipomoea* representing Convolvulaceae, the sister family of Solanaceae but it appears to be functional in *Coffea* of Rubiaceae [75] used as a distant outgroup of Lamiales. The *infA* gene seem to have become non-functional in the ancestor of Solanales multiple times independently. In Solanaceae the pseudogenization further continued with a monophyletic 124-bp deletion in the ancestor of the genus *Solanum*. Further changes appeared in four protein-coding genes; there is a 64-bp deletion in *psbD* of *Iochroma tingoanum* while 31-bp was deleted from the *rpl20* gene in members of Physaleae. *Capsicum lycianthoides* Bitter had a unique 15-bp insertion in the *rpl33* gene. The *accD* gene, which encodes one of the four subunits of the acetyl-CoA carboxylase enzyme in most chloroplasts show a 24-bp insertion in the members of the 'x = 12 clade' [61]. This seems to be an ancestral trait shared by members of Nicotianoideae and Solanoideae and maintained in *Datura* L., *Nicotiana*, *Physalis* L. and *Iochromas* but lost independently in *Hyoscyamus* L., *Capsicum* L. and *Solanum*. The latter two went through a characteristic 141-bp and a small 9-bp insertion. The 141-bp deletion was also confirmed in *Capsicum* by Jo et al. [72]. The small plastid RNA (*sprA*) gene, which includes a complementary segment to the pre-16S rRNA shows high variability among Solanaceae. Functional *sprA* copies were present in most Solanaceae but several mutation event indicate it has been non-functional in some groups. A 52-bp deletion appeared in *Capsicum* at the 5' and further 37-bp were deleted in *Iochromas* while *Physalis* showed an autapomorphic 14-bp insertion (S3 Fig). The function *sprA* has been lost independently multiple times once in *Iochrominae* and in *Capsaceae*, however, the gene remained functional in *Capsicum lycianthoides*.

Genomic changes also affect tRNA genes and neighboring regions. The most notable change is the duplication of the original phenylalanine (*trnF*-GAA) gene in a tandem array composed by multiple pseudogene copies in Solanaceae. The pseudogene copies are composed of several highly structured motifs that are partial residues or entire parts of the anticodon, T- and D-domains of the original *trnF* gene [66]. Previously it was shown that these copies are subjected to possible inter- or intrachromosomal recombination events [67] and they have high taxonomic relevance uniting a unique plastid clade of Pseudosolanoids [68]. They provide support for previous results [10, 59] separating the Atropina and Juanulloae clades from Solanaceae, Capsaceae, Physaleae, Datureae and Salpichroina [68]. Another tRNA related structural change is apparent in the group II intron of *trnA*-UGC, where 108-bp was deleted in *Nicotiana* and extended up to 147-bp in *Atropa* L. and *Hyoscyamus*.

Gene expression analyses

We carried out the expression analysis of 85 protein-coding genes (Table 3). As we were mostly interested about CDS/gene features we used only these annotation types for read mapping. We also used the RNA-seq data set to verify start/stop codon positions and further ultimate or penultimate editing sites from the reannotation process. A total of 147,721 reads were mapped to the bittersweet plastid genome with an average 112× read depth. The largest portion of reads 25,910 (17.53%) and 12,582 (8.51%) was derived from adenosine triphosphate

Table 3. RNA Expression of protein-coding genes in the *Solanum dulcamara* chloroplast genome. Reads per kilobase per million (RPKM), fragments per kilobase of exon per million fragments mapped (FPKM) and transcripts per million (TPM) for transcript variants.

Gene	Location min.	Max	Length	FPKM	RPKM	TPM
<i>atpB</i>	54,285	55,781	1,497	278926.7	278926.7	232422.2
<i>atpE</i>	53,887	54,288	402	238932.2	238932.2	199095.9
<i>clpP</i>	71,842	73,864	591	120109.6	120109.6	100084.1
<i>rps7</i>	142,238	142,705	468	91956.3	91956.3	76624.7
<i>rps7</i>	98,701	99,168	468	88572.2	88572.2	73804.8
<i>psbM</i>	30,605	30,709	105	22431.7	22431.7	18691.7
<i>psbA</i>	552	1,613	1,062	21738.5	21738.5	18114.1
<i>ycf1</i>	125,388	131,069	5,682	21573.2	21573.2	17976.3
<i>psbK</i>	7,750	7,935	186	21287.0	21287.0	17737.9
<i>psaJ</i>	68,897	69,031	135	19101.3	19101.3	15916.6
<i>rbcL</i>	56,597	58,030	1,434	13932.8	13932.8	11609.9
<i>rpl20</i>	70,391	70,777	387	11700.0	11700.0	9749.3
<i>psbI</i>	8,248	8,406	159	11493.2	11493.2	9576.9
<i>rps12</i>	71,590	142,184	372	11407.7	11407.7	9505.7
<i>rps12</i>	71,590	100,015	372	11243.9	11243.9	9369.3
<i>psbJ</i>	65,856	65,978	123	10565.0	10565.0	8803.5
<i>atpF</i>	11,989	13,234	555	8579.1	8579.1	7148.8
<i>psbE</i>	66,378	66,629	252	8540.8	8540.8	7116.8
<i>rps16</i>	5,077	6,199	267	7528.7	7528.7	6273.4
<i>atpH</i>	13,637	13,882	246	7180.9	7180.9	5983.6
<i>ycf1</i>	110,382	111,527	1,146	6963.1	6963.1	5802.2
<i>rps18</i>	69,855	70,160	306	6768.2	6768.2	5639.8
<i>rps15</i>	124,723	124,986	264	6537.5	6537.5	5447.5
<i>rps19</i>	85,655	85,933	279	6258.8	6258.8	5215.3
<i>rpl22</i>	85,135	85,602	468	6225.9	6225.9	5187.8
<i>rps14</i>	38,024	38,326	303	6098.1	6098.1	5081.4
<i>ndhH</i>	123,425	124,606	1,182	5531.4	5531.4	4609.1
<i>psbT</i>	76,034	76,138	105	5511.2	5511.2	4592.4
<i>rpl16</i>	82,913	84,349	405	5113.7	5113.7	4261.1
<i>psbZ</i>	37,053	37,241	189	4941.9	4941.9	4117.9
<i>psaC</i>	118,619	118,864	246	4704.7	4704.7	3920.3
<i>cemA</i>	62,915	63,604	690	4472.9	4472.9	3727.1
<i>rps3</i>	84,494	85,150	657	4249.4	4249.4	3540.9
<i>ycf3</i>	43,702	45,689	507	4245.1	4245.1	3537.4
<i>psbC</i>	34,984	36,369	1,386	4211.8	4211.8	3509.6
<i>psbB</i>	74,308	75,834	1,527	4135.4	4135.4	3445.9
<i>rpl33</i>	69,463	69,663	201	3838.7	3838.7	3198.7
<i>ndhA</i>	121,171	123,423	1,092	3830.3	3830.3	3191.7
<i>rpl2</i>	153,916	155,406	825	3704.0	3704.0	3086.5
<i>psaB</i>	38,445	40,649	2,205	3480.8	3480.8	2900.4
<i>petN</i>	29,403	29,492	90	3384.1	3384.1	2819.9
<i>psaA</i>	40,675	42,927	2,253	3343.5	3343.5	2786.1
<i>rpl2</i>	86,000	87,490	825	3322.6	3322.6	2768.6
<i>psbD</i>	33,939	35,000	1,062	3259.8	3259.8	2716.3
<i>petB</i>	76,806	78,207	652	3238.8	3238.8	2698.8
<i>ndhK</i>	50,792	51,535	744	3097.5	3097.5	2581.1

(Continued)

Table 3. (Continued)

Gene	Location min.	Max	Length	FPKM	RPKM	TPM
<i>ndhI</i>	120,574	121,077	504	2860.4	2860.4	2383.5
<i>ndhB</i>	96,202	98,413	1,533	2834.4	2834.4	2361.9
<i>ndhB</i>	142,993	145,204	1,533	2794.7	2794.7	2328.7
<i>rps2</i>	16,048	16,758	711	2770.1	2770.1	2308.3
<i>atpA</i>	10,411	11,934	1,524	2618.0	2618.0	2181.5
<i>atpI</i>	15,056	15,799	744	2565.4	2565.4	2137.6
<i>rps8</i>	81,838	82,242	405	2506.7	2506.7	2088.8
<i>rpl14</i>	82,410	82,778	369	2366.1	2366.1	1971.6
<i>ndhJ</i>	50,210	50,686	477	2256.1	2256.1	1879.9
<i>psbN</i>	76,212	76,343	132	2230.4	2230.4	1858.6
<i>ndhC</i>	51,526	51,888	363	2097.6	2097.6	1747.9
<i>petD</i>	78,398	79,611	483	1639.5	1639.5	1366.2
<i>psbH</i>	76,455	76,676	222	1554.9	1554.9	1295.6
<i>ndhG</i>	119,645	120,175	531	1453.1	1453.1	1210.8
<i>matK</i>	2,136	3,665	1,530	1446.5	1446.5	1205.4
<i>petG</i>	67,909	68,022	114	1424.9	1424.9	1187.3
<i>rpoC1</i>	21,302	24,105	2,067	1414.5	1414.5	1178.7
<i>rps11</i>	80,882	81,298	417	1412.1	1412.1	1176.6
<i>petA</i>	63,824	64,786	963	1244.0	1244.0	1036.6
<i>rpoA</i>	79,803	80,816	1,014	1201.5	1201.5	1001.1
<i>ycf4</i>	61,594	62,148	555	1170.7	1170.7	975.5
<i>ndhE</i>	119,116	119,421	306	1128.0	1128.0	940.0
<i>rpl23</i>	153,616	153,897	282	1116.0	1116.0	930.0
<i>psaI</i>	61,037	61,147	111	1097.5	1097.5	914.6
<i>rpl23</i>	87,509	87,790	282	1080.0	1080.0	900.0
<i>rpl32</i>	114,524	114,691	168	966.9	966.9	805.7
<i>accD</i>	58,765	60,288	1,524	906.0	906.0	754.9
<i>rps4</i>	46,706	47,311	606	770.6	770.6	642.1
<i>rpoB</i>	24,111	27,338	3,228	673.0	673.0	560.8
<i>petL</i>	67,627	67,722	96	634.5	634.5	528.7
<i>ndhD</i>	116,999	118,501	1,503	580.9	580.9	484.1
<i>rpoC2</i>	16,983	21,149	4,167	438.5	438.5	365.4
<i>rpl36</i>	81,400	81,513	114	356.2	356.2	296.8
<i>ycf2</i>	88,118	94,960	6,843	308.6	308.6	257.1
<i>ycf2</i>	146,446	153,288	6,843	281.9	281.9	234.9
<i>ndhF</i>	111,507	113,729	2,223	246.6	246.6	205.5
<i>ccsA</i>	115,826	116,767	942	215.5	215.5	179.6
<i>ycf15</i>	95,045	95,308	264	76.9	76.9	64.1
<i>ycf15</i>	146,098	146,361	264	76.9	76.9	64.1

<https://doi.org/10.1371/journal.pone.0196069.t003>

(ATP) synthase genes and from the photosystem II (PSII) complex. All genes were normally expressed while the five most abundant were *atpB*, *atpE*, *clpP*, *rps7* and *psbM* (>10,000 FPKM). The assembled consensus sequence from the mapped reads (148,110 bp long) covered 95.22% of the genome spanning through also intergenic spacer (IGS) sequences. Accordingly, a nearly complete pseudo *Solanum dulcamara* plastid genome was unexpectedly obtained by means of transcriptome data. We found multiple transcripts mapping to several non-functional genes for example *ycf15*, *infA*, or to truncated pseudogenes *ψycf1* and *ψrps19* at the JLA

(IRa/LSC). From these *infA*, *ψycf1* and *ψrps19* were nearly completely covered (S4 Fig) showing that they are indeed transcribed, while *ycf15* had sparse coverage. This indicates that transcriptome sequencing captured both primary and processed mRNA sequences of the plastome. The detected and mapped reads of the bittersweet plastid RNA population could be grouped into three major types i) mRNAs ii) non-coding RNAs from IGS regions and iii) transcribed non-coding RNAs (rRNAs and tRNAs). Similar patterns were observed by Shi et al. [76] and also in earlier studies using Northern blot hybridization where 90% of the plastid genome was found to be transcribed [77]. Such patterns could be caused by transcriptional uncoupling of genes in polycistronic clusters [78]. Non-coding RNAs (ncRNAs) in the plastome are further transcribed from intergenic regions (IGSs), which play important role in post-transcriptional regulation [79]. Cyanobacteria contain several ncRNAs making it plausible that also plastomes harbor a wide variety of undetected regulatory ncRNAs [80]. These results show that non-functional genes are transcribed as a precursor polycistronic transcript, which are later edited during pre-mRNA maturation. In order to activate the function of other genes plastid primary transcripts are edited and expression in the plastome mainly occurs at a post-transcriptional stage. The multiple transcription arrangement leading to the full transcription of plastid genomes is a prokaryotic ancestral trait still preserved in eukaryotic cells billion years after the primary endosymbiosis [81, 82].

Plastid RNA editing

Chloroplast RNA editing was first discovered in 1991 [83] and it could be defined as the post-transcriptional modification of pre-RNAs by insertion, deletion or substitution of specific nucleotides to form functional RNAs. In the plastid genome this processing machinery is crucial to alter the long pre-RNA transcripts as detailed above. The most frequent editing events in plants are C-to-U changes, however, U-to-C editing has also been observed [84]. RNA editing is absent in liverworts and green algae while it is abundant in lycophytes, ferns and hornworts [85]. To gain insight to the RNA metabolism of bittersweet we first predicted 28 RNA editing sites out of 35 plastid genes with PREP (Table 4). We aligned RNA read sequences using bittersweet as a reference genome and by variant searching we confirmed 23 editing sites from those predicted with PREP. We found four additional editing sites with variant search not detected by PREP resulting in 27 confirmed editing sites. From these 25 (92.5%) were C-to-U changes and two were A-to-G and G-to-U conversions resulting in non-synonymous amino acid changes. The percentage of conversion rates for each edit varied between 25 to 95.9% according to the calculated ratio between the numbers of reads with an alternate base compared with the reference. Some edits showed high rates (>90%) for *atpF*, *ndhB*, *petB*, *psbE* and *rps14* genes making it clear that these forms are highly abundant among processed RNAs in bittersweet. Edits of these particular genes has also been reported in previous studies of embryophytes [86, 87] suggesting the conserved feature of such sites. It has been proposed that RNA editing is of monophyletic origin and evolved as a mechanism to conserve certain codons [88]. For example the start codon (AUG) of the *psbL* and *ndhD* is RNA edited (C-to-U) in all Solanaceae except in *Datura stramonium* where the start codon of *psbL* remains unedited.

Conclusions

Comparison of chloroplast genome organization not only provide us with valuable information for understanding the processes of chloroplast evolution, but also gives insights into the mechanisms underlying genomic rearrangements [25]. Furthermore, investigation of plastid genome structures could trigger further breakthroughs in applied sciences. For example herbicides like PSI and PSII inhibitors have their target genes in the chloroplast genome thus

Table 4. RNA editing sites in the *Solanum dulcamara* chloroplast genome.

Gene Name	Length	Strand	Region	Nt pos	AA pos	Effect	Nt Change	Score	RNASeq	PREP	Number of reads
<i>atpF</i>	1246	+	LSC	92	31	CCA (P) => CUA (L)	C => U	0.86	+	+	U; 49 (90.7%), C; 5 (9.3%)
<i>ndhA</i>	2258	+	SSC	341	114	UCA (S) => UUA (L)	C => U	1	+	+	U; 35 (70%), C; 15 (30%)
<i>ndhA</i>	2258	+	SSC	566	189	UCA (S) => UUA (L)	C => U	1	+	+	U; 20 (34.4%), C; 38 (65.6%)
<i>ndhA</i>	2258	+	SSC	1073	358	UCC (S) => UUC (F)	C => U	1	+	+	U; 49 (74.2%), C; 17 (25.8%)
<i>ndhB</i>	2212	+	IR	149	50	UCA (S) => UUA (L)	C => U	1	+	+	U; 33 (86.8%), C; 5 (13.1%)
<i>ndhB</i>	2212	+	IR	467	156	CCA (P) => CUA (L)	C => U	1	+	+	U; 34 (87.1%), C; 5 (12.9%)
<i>ndhB</i>	2212	+	IR	586	196	CAU (H) => UAU (Y)	C => U	1	+	+	U; 26 (82.3%), C; 8 (17.7%)
<i>ndhB</i>	2212	+	IR	611	204	UCA (S) => UUA (L)	C => U	0.80	+	+	U; 33 (89.1%), C; 4 (10.9%)
<i>ndhB</i>	2212	+	IR	737	246	CCA (P) => CUA (L)	C => U	1	+	+	U; 47 (95.9%), C; 2 (4.1%)
<i>ndhB</i>	2212	+	IR	746	249	UCU (S) => UUU (F)	G => U	1	+	+	U; 40 (95.2%), C; 2 (4.8%)
<i>ndhB</i>	2212	+	IR	780	260	UGG (P) => UGU (C)	C => U	-	+	-	U; 32 (50.8%), G; 31 (49.2%)
<i>ndhB</i>	2212	+	IR	830	277	UCA (S) => UUA (L)	C => U	1	+	+	U; 44 (97.1%), C; 1 (2.9%)
<i>ndhB</i>	2212	+	IR	836	279	UCA (S) => UUA (L)	C => U	1	-	+	-
<i>ndhB</i>	2212	+	IR	1481	494	CCA (P) => CUA (L)	C => U	1	+	+	U; 20 (52.6%), C; 18 (47.4%)
<i>ndhD</i>	1504	+	SSC	2	1	ACG (T) => AUG (M)	C => U	-	+	-	U; 40 (95.2%), C; 2 (4.8%)
<i>ndhF</i>	2223	+	SSC	290	97	UCA (S) => UUA (L)	C => U	1	-	+	-
<i>petB</i>	1398	-	LSC	1168	390	CGG (R) => UGG (W)	C => U	1	+	+	U; 15 (93.8%), C; 1 (6.2%)
<i>petB</i>	1398	-	LSC	1361	454	CCA (P) => CUA (L)	C => U	1	+	+	U; 23 (74.2%), C; 8 (25.8%)
<i>psbE</i>	252	+	LSC	214	72	CCU (P) => UCU (S)	C => U	1	+	+	U; 112 (93.3%), C; 8 (6.7%)
<i>psbL</i>	124	+	LSC	2	1	ACG (T) => AUG (M)	C => U	-	+	-	U; 40 (95.2%), C; 2 (4.8%)
<i>rpl20</i>	387	+	LSC	308	103	UCA (S) => UUA (L)	C => U	0.86	+	+	U; 107 (56.6%), C; 82 (43.4%)
<i>rpoA</i>	1014	+	LSC	830	277	UCA (S) => UUA (L)	C => U	1	+	+	U; 8 (61.5%), C; 5 (38.5%)
<i>rpoA</i>	1014	+	LSC	903	301	AUG (M) => GUG (V)	A => G	-	+	-	G; 25 (62.5%), A; 15 (37.5%)
<i>rpoB</i>	3213	+	LSC	338	113	UCU (S) => UUU (F)	C => U	1	+	+	U; 15 (75%), C; 5 (25%)
<i>rpoB</i>	3213	+	LSC	473	158	UCA (S) => UUA (L)	C => U	0.86	+	+	U; 13 (76.5%), C; 4 (23.5%)
<i>rpoB</i>	3213	+	LSC	551	184	UCA (S) => UUA (L)	C => U	1	-	+	-
<i>rpoB</i>	3213	+	LSC	2000	667	UCU (S) => UUU (F)	C => U	1	-	+	-
<i>rpoB</i>	3213	+	LSC	2426	809	UCA (S) => UUA (L)	C => U	0.86	+	+	U; 5 (25%), C; 15 (75%)
<i>rpoC1</i>	2783	+	LSC	41	14	UCA (S) => UUA (L)	C => U	1	+	+	U; 5 (27.7%), C; 13 (72.3%)
<i>rpoC2</i>	4167	+	LSC	119	40	CCC (P) => CUC (L)	C => U	-	+	-	U; 10 (37.1%), C; 17 (62.9%)
<i>rpoC2</i>	4167	+	LSC	3731	1244	UCA (S) => UUA (L)	C => U	0.86	-	+	-
<i>rps2</i>	711	+	LSC	134	45	ACA (T) => AUA (I)	C => U	-	+	-	C; 8 (38.1%), U; 13 (61.9%)
<i>rps2</i>	711	+	LSC	248	83	UCA (S) => UUA (L)	C => U	1	+	+	C; 5 (31.3%), U; 11 (68.7%)
<i>rps14</i>	303	+	LSC	80	27	UCA (S) => UUA (L)	C => U	1	+	+	C; 5 (5.8%), U; 81 (94.2%)

<https://doi.org/10.1371/journal.pone.0196069.t004>

understanding the chloroplast genome may indirectly support the exploration of herbicide resistance and development of novel control methods [89]; while plastid engineering can also be useful to develop resistance to various abiotic and biotic stress factors based on discovered resistance traits. Here we report the complete chloroplast genome sequence of *Solanum dulcamara* as a genomic tool for potential plastid genome comparative studies. We also present the reannotation of Solanaceae plastid genomes using manual curation using *S. dulcamara* as a reference. Based on the reannotated genome sequences we introduce a hypothesis of the ancestral plastid genome organization of Solanaceae and the rearrangements unique to some major clades. The ancestral plastid genome of Solanaceae had two degraded non-functional genes, *infA* and truncated *ycf1* copy, a deletion in the *trnA* intron and the appearance of a highly divergent gene (*sprA*). Our ancestral genome reconstruction suggests further rearrangements in the stem branch of Solanoideae by the expansion of the IR and the occurrence of a truncated

$\psi rps19$ copy at the JLA as a consequence of the expansion. This has been followed by independent rearrangements in deeper nodes such as the accumulation of *trnF* pseudogenes in tandem arrays at a clade referred to as the 'Pseudosolanoids' [68] or by the pseudogenization of *sprA* in Physaleae and Capsiceae by two deletions. Further degradation of the *infA* pseudogene is specific for the largest genus *Solanum*, including tomato and potato.

Supporting information

S1 Data. MAFFT sequence alignment for 35 complete plastid genome sequences used in phylogenetic analysis.

(RAR)

S2 Data. NEXUS file containing the binary coding used to map genomic changes appearing in the chloroplast genome.

(RAR)

S3 Data. Annotated checklist of SSRs in *Solanum dulcamara* plastid genome hits founds by MISA.

(RAR)

S4 Data. Reannotation file of Solanaceae plastid genomes in Geneious format, accessible with 7.1 or later version.

(RAR)

S5 Data. Reannotation files in GFF and GB file format.

(ZIP)

S1 Table. NCBI GenBank accession numbers used in this study.

(DOCX)

S2 Table. List of genes in the chloroplast genome of bittersweet.

(DOCX)

S3 Table. The genes having intron in the *Solanum dulcamara* plastid genome and the length of the exons and introns.

(DOCX)

S4 Table. Comparison of major features of *Solanum dulcamara* and nine Solanaceae plastid genomes.

(DOCX)

S5 Table. Relative synonymous codon usage (RSCU) of *Solanum dulcamara* is given in parentheses following the codon frequency.

(DOCX)

S6 Table. Estimates of average evolutionary divergence over 80 protein coding-gene sequences from Solanaceae.

(DOCX)

S7 Table. Total number of perfect simple sequence repeats (SSRs) identified within the chloroplast genome of *Solanum dulcamara*.

(DOCX)

S8 Table. List of annotation errors found in Solanaceae chloroplast genomes.

(XLSX)

S1 Fig. Best scoring maximum likelihood trees obtained with RAxML and the most parsimonious tree generated with TNT.

(DOCX)

S2 Fig. Visualization alignment of chloroplast genome sequences with mVISTA-based identity plots.

(PNG)

S3 Fig. Alignment of the *sprA* gene in Solanaceae.

(PDF)

S4 Fig. RNAseq reads mapped to the genomic region of *ycf15* pseudogene.

(PDF)

Acknowledgments

We thank staff and colleagues of the Viikki Biocenter who kindly contributed reagents, materials and analyses tools for our study.

Author Contributions

Conceptualization: Péter Poczai.

Data curation: Péter Poczai.

Formal analysis: Ali Amiryousefi, Jaakko Hyvönen, Péter Poczai.

Investigation: Péter Poczai.

Methodology: Ali Amiryousefi, Péter Poczai.

Project administration: Péter Poczai.

Resources: Jaakko Hyvönen, Péter Poczai.

Software: Ali Amiryousefi, Péter Poczai.

Supervision: Jaakko Hyvönen, Péter Poczai.

Validation: Jaakko Hyvönen, Péter Poczai.

Visualization: Ali Amiryousefi, Jaakko Hyvönen, Péter Poczai.

Writing – original draft: Péter Poczai.

Writing – review & editing: Jaakko Hyvönen, Péter Poczai.

References

1. Knapp S. The revision of the Dulcamaroid clade of *Solanum* L. (Solanaceae). *PhytoKeys*. 2013; 22: 1–432.
2. Máthé I, Máthé I Jr. Variations in alkaloids in *Solanum dulcamara* L. In: Hawkes JG, Lester RN, Skelding AD (Eds) *The biology and taxonomy of the Solanaceae*. Academic Press, London, 1979; pp. 211–222.
3. Kumar P, Sharma B, Bakshi N. Biological activity of alkaloids from *Solanum dulcamara* L. *Nat Prod Res*. 2009; 23: 719–723. <https://doi.org/10.1080/14786410802267692> PMID: 19418354
4. D'Arcy WG. Solanaceae studies II: Typification of subdivisions of *Solanum*. *Ann Miss Bot Gard*. 1972; 59: 262–278.
5. Nee M. Synopsis of *Solanum* in the New World. In: Nee M, Symon DE, Lester RN, Jessop JP (Eds) *Solanaceae IV: Advances in biology and utilization*. Royal Botanic Gardens, Kew, 1999; 285–333.

6. Lester RN. Evolutionary relationships of tomato, potato, pepino and wild species of *Lycopersicon* and *Solanum*. In: Hawkes JG, Lester RN, Nee M and Estrada-R N (eds.), *Solanaceae III: Taxonomy, Chemistry and Evolution*. Roy. Bot. Gardens, Kew. 1991; pp. 283–301
7. Child A, Lester RN. Synopsis of the genus *Solanum* L. and its infrageneric taxa. In: van den Berg RG, Barendse GWM, van der Weerden GM, Mariani C (eds) *Solanaceae V: advances in taxonomy and utilization*. Nijmegen University Press, 2001; pp 39–52.
8. Bohs L. Major clades in *Solanum* based on *ndhF* sequence data. In: Keating RC, Hollowell VC, Croat TB (eds) *A Festschrift for William G. D'Arcy: the legacy of a taxonomist*. Missouri Botanical Garden Press, St. Louis (Monographs in systematic botany from the Missouri Botanical Garden. 2005; 104: 27–49)
9. Weese T, Bohs L. A three-gene phylogeny of the genus *Solanum* (Solanaceae). *Syst Bot.* 2007; 32: 445–463.
10. Särkinen T, Bohs L, Olmstead RG, Knapp S. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evol Biol.* 2013; 13: 214. <https://doi.org/10.1186/1471-2148-13-214> PMID: 24283922
11. Särkinen T, Barboza GE, Knapp. True back nightshades: phylogeny and delimitation of the Morelloid clade of *Solanum*. *Taxon.* 2015; 64:945–958.
12. Takács AP, Kazinczi G, Horváth J, Pribék D. New host-virus relations between different *Solanum* species and viruses. *Meded Rijkuniv Gent Fak Landbouwkd Toegep Biol Wt.* 2001; 66: 183–186.
13. Perry KL, McLane H. Potato virus m in bittersweet nightshade (*Solanum dulcamara*) in New York State. *Plant Dis.* 2011; 95: 619–623.
14. Hajianfar R, Kolics B, Cernák I, Wolf I, Polgár Zs, Taller J. Expression of biotic stress response genes to *Phytophthora infestans* inoculation in White Lady, a potato cultivar with race-specific resistance to late blight. *Physiol Mol Plant Pathol.* 2016; 93:22–28.
15. Golas TM, Weerden GMVD, Berg RGVD, Mariani C, Allefs JJHM. Role of *Solanum dulcamara* L. in potato late blight epidemiology. *Potato Res.* 2010; 53: 69–81.
16. Golas TM, Feron RMC, van den Berg RG, van der Weerden GM, Mariani C, Allefs JJHM Genetic structure of European accessions of *Solanum dulcamara* L. (Solanaceae). *Plant Syst Evol.* 2010a; 285: 103–110.
17. Poczai P, Varga I, Bell NE, Hyvönen J. Genetic diversity assessment of bittersweet (*Solanum dulcamara*, Solanaceae) germplasm using conserved DNA-derived polymorphism and intron-targeting markers. *Ann Appl Biol.* 2011; 159: 141–153.
18. Vallejo-Marín M, O'Brien HE. Correlated evolution of self-incompatibility and clonal reproduction in *Solanum* (Solanaceae). *New Phytol.* 2006; 173:415–421.
19. Hewitt GM. Genetic consequences of climatic oscillations in the Quaternary. *Phil Trans R Soc Lond B.* 2004; 359: 183–195.
20. Oldenburg DJ, Bendich AJ. DNA maintenance in plastids and mitochondria of plants. *Frontiers Plant Sci.* 2015; 6:883.
21. Oldenburg DJ, Bendich AJ. The linear plastid chromosomes of maze: terminal sequences, structures, and implications for DNA replication. *Curr Genet.* 2016; 62:431–442. <https://doi.org/10.1007/s00294-015-0548-0> PMID: 26650613
22. Daniell H, Lin C-S, Yu M, Chang W-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 2016; 17:134. <https://doi.org/10.1186/s13059-016-1004-2> PMID: 27339192
23. Kim KJ, Lee HL. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* 2004; 11: 247–261. PMID: 15500250
24. Weng M-L, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol Biol Evol.* 2014; 31:645–659. <https://doi.org/10.1093/molbev/mst257> PMID: 24336877
25. Poczai P, Hyvönen J. The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, Bromeliaceae) and its comparative analysis. *PloS ONE.* 2017; 12: e0187199. <https://doi.org/10.1371/journal.pone.0187199> PMID: 29095905
26. Karol KG, Arumuganathan K, Boore JL, Duffy AM, Everett KDE, Hall JD et al. Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evol Biol.* 10:321. <https://doi.org/10.1186/1471-2148-10-321> PMID: 20969798
27. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale

- evolutionary patterns. *Proc Natl Acad Sci USA*. 2007; 104:19369–19374. <https://doi.org/10.1073/pnas.0709121104> PMID: 18048330
28. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE* 2012; 7:e31468. <https://doi.org/10.1371/journal.pone.0031468> PMID: 22384027
29. Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Meth*. 2010; 6:22.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> PMID: 24695404
31. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012; 28: 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199> PMID: 22543367
32. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using Bruijn graphs. *Genome Res*. 2008; 18: 821–829. <https://doi.org/10.1101/gr.074492.107> PMID: 18349386
33. Moore MJ, Bell CD, Soltis PS, Soltis DE. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci USA*. 2007; 104: 19363–19368. <https://doi.org/10.1073/pnas.0708072104> PMID: 18048334
34. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*. 2004; 20: 3252–3255. <https://doi.org/10.1093/bioinformatics/bth352> PMID: 15180927
35. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acid Res*. 1997; 25: 955–964. PMID: 9023104
36. Liu C, Shi L, Chen H, Zhang J, Lin X, Guan X. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*. 2012; 13: 715. <https://doi.org/10.1186/1471-2164-13-715> PMID: 23256920
37. McKain MR, Hartsock RH, Wohl MM, Kellogg EA. Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. *Bioinformatics*. 2017; 33: 130–132. <https://doi.org/10.1093/bioinformatics/btw583> PMID: 27634949
38. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R et al. GeSeq—versatile and accurate annotation of organelle genomes. *Nucl Acids Res*. 2017; 45 (W1): W6–W11. <https://doi.org/10.1093/nar/gkx391> PMID: 28486635
39. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW)—a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet*. 2007; 52: 267–274. <https://doi.org/10.1007/s00294-007-0161-y> PMID: 17957369
40. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016; 33: 1870–1874. <https://doi.org/10.1093/molbev/msw054> PMID: 27004904
41. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acid Res*. 2004; 32: W273–279. <https://doi.org/10.1093/nar/gkh458> PMID: 15215394
42. Amiryousefi A, Hyvönen J, Poczar P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*. 2018; bty220, <https://doi.org/10.1093/bioinformatics/bty220>
43. Kurtz S, Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*. 1999; 15: 426–427. PMID: 10366664
44. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.) *Theor Appl Genet*. 2003; 106: 411–422. <https://doi.org/10.1007/s00122-002-1031-0> PMID: 12589540
45. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009; 10: R25. <https://doi.org/10.1186/gb-2009-10-3-r25> PMID: 19261174
46. Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acid Res*. 2009; 37: W253–W259. <https://doi.org/10.1093/nar/gkp337> PMID: 19433507
47. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012; arXiv preprint arXiv:1207.3907 [q-bio.GN]
48. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol Biol Evol*. 2013; 30: 772–780. <https://doi.org/10.1093/molbev/mst010> PMID: 23329690

49. RAxML Next Generation: faster, easier-to-use and more flexible. 2018; <https://doi.org/10.5281/zenodo.593079>
50. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol.* 2017; 34:772–773. <https://doi.org/10.1093/molbev/msw260> PMID: 28013191
51. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and high-performance computing. *Nature Math.* 2012; 9:772.
52. Nguyen L-T, Schmidt HA, Haeseler VA, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015; 32:268–274. <https://doi.org/10.1093/molbev/msu300> PMID: 25371430
53. Goloboff PA, Farris JS, Nixon KC. TNT, a free program for phylogenetic analysis. *Cladistics.* 2008; 24:774–786.
54. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. 2017; Version 3.2 <http://mesquiteproject.org>
55. Stöver BC, Müller KF. TreeGraph2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinf.* 2010; 11:7.
56. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 1986; 5:2043–2049. PMID: 16453699
57. Vera A, Sugiura M. A novel RNA gene in the tobacco plastid genome: its possible role in the maturation of 16S rRNA. *EMBO J.* 1994; 13:2211–2217. PMID: 7514532
58. Sugita M, Svab Z, Maliga P, Sugiura M. Targeted deletion of *sprA* from the tobacco plastid genome indicates that the encoded small RNA is not essential for pre-16S rRNA maturation in plastids. *Mol Gen Genet.* 1997; 257:23–27. PMID: 9439565
59. Olmstead RG, Bohs L, Migid HA, Santiago-Valentin E, Garcia VF, Collier SM. A molecular phylogeny of the Solanaceae. *Taxon.* 2008; 57:1159–1181.
60. Smith SD, Baum DA. Phylogenetics of the florally diverse Andean clade Iochrominae (Solanaceae). *Am J Bot.* 2006; 93:1140–1153. <https://doi.org/10.3732/ajb.93.8.1140> PMID: 21642180
61. Olmstead RG, Palmer JD. A chloroplast DNA phylogeny of the Solanaceae: subfamilial relationships and character evolution. *Ann Miss Bot Gard.* 1992; 79:346–360.
62. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature.* 2012; 485:635–641. <https://doi.org/10.1038/nature11119> PMID: 22660326
63. Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry S et al. Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nature Plants.* 2016; 2:16074. <https://doi.org/10.1038/nplants.2016.74> PMID: 27255838
64. Olmstead RG, Palmer JD. A chloroplast DNA phylogeny of the Solanaceae: subfamilial relationships and character evolution. *Ann. Miss. Bot. Gard.* 1992; 79:346–360.
65. Chung H-J, Jung JD, Park H-W, Kim J-H, Cha HW, Min SR et al. The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. *Plant Cell Rep.* 2006; 25:1369–1379. <https://doi.org/10.1007/s00299-006-0196-4> PMID: 16835751
66. Poczai P, Hyvönen J. Identification and characterization of plastid *trnF*(GAA) pseudogenes in four species of *Solanum* (Solanaceae). *Biotech Lett.* 2011; 33:2317–2323.
67. Poczai P, Hyvönen J. Plastid *trnF* pseudogenes are present in *Jalotmata*, the sister genus of *Solanum* (Solanaceae): molecular evolution of tandemly repeated structural mutations. *Gene.* 2013; 530:143–150. <https://doi.org/10.1016/j.gene.2013.08.013> PMID: 23962687
68. Poczai P, Hyvönen J. Discovery of novel plastid phenylalanine (*trnF*) pseudogenes defines a distinctive clade in Solanaceae. *SpringerPlus.* 2013; 2:459. <https://doi.org/10.1186/2193-1801-2-459> PMID: 24083106
69. Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM. The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. *Mol Biol Evol.* 2002; 19:1602–1612. <https://doi.org/10.1093/oxfordjournals.molbev.a004222> PMID: 12200487
70. Kahlau S, Aspinall S, Gray JC, Bock R. Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. *J Mol Evol.* 2006; 63:194–207. <https://doi.org/10.1007/s00239-005-0254-5> PMID: 16830097
71. Daniell H, Lee S-B, Grevich J, Saski C, Quesada-Vargas T, Guda C et al. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other

- Solanaceae genomes. *Theor Appl Genet.* 2006; 112: 1503–1518. <https://doi.org/10.1007/s00122-006-0254-x> PMID: 16575560
72. Jo YD, Park J, Kim J, Song W, Hur C-G, Lee Y-H et al. Complete sequencing and comparative analyses of the pepper (*Capsicum annuum* L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. *Plant Cell Rep.* 2011; 30: 217–229. <https://doi.org/10.1007/s00299-010-0929-2> PMID: 20978766
73. Sanchez-Puerta MV, Abbona CC. The chloroplast genome of *Hyoscyamus niger* and a phylogenetic study of the tribe Hyoscyameae (Solanaceae). *PLoS ONE.* 2014; 9: e98353. <https://doi.org/10.1371/journal.pone.0098353> PMID: 24851862
74. Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L, Yitao W. Complete chloroplast genome sequence of poisonous and medicinal plant *Datura stramonium*: organizations and implications for genetic engineering. *PLoS ONE.* 2014; 9: e110656. <https://doi.org/10.1371/journal.pone.0110656> PMID: 25365514
75. Samson N, Bausher MG, Lee S-B, Jansen RK, Daniell H. The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Plant Biotech. J.* 2007; 5:339–353.
76. Shi C, Wang S, Xia E-H, Jiang J-J, Zeng F-C, Gao L-Z. Full transcription of the chloroplast genome in photosynthetic eukaryotes. *Sci Rep.* 2016; 6:30135. <https://doi.org/10.1038/srep30135> PMID: 27456469
77. Woodbury NW, Roberts LL, Palmer JD, Thompson WF. A transcription map of the pea chloroplast genome. *Curr Genet.* 1988; 14: 75–89.
78. Zhelyazkova P, Sharma CM, Förstner KU, Liere K, Vogel J, Börner T. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plan Cell.* 2012; 24: 1123–1136.
79. Germain A, Hotto AM, Barkan A, Stern DB. RNA processing and decay in plastids. *WIREs RNA.* 2013; 4: 295–316. <https://doi.org/10.1002/wrna.1161> PMID: 23536311
80. Hotto AM, Germain A, Stern DB. Plastid non-coding RNAs: emerging candidates for gene regulation. *Trends Plant Sci.* 2012; 17: 737–744. <https://doi.org/10.1016/j.tplants.2012.08.002> PMID: 22981395
81. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Rev Genet.* 2009; 10: 833–844. <https://doi.org/10.1038/nrg2683> PMID: 19920851
82. Shi C, Liu Y, Huang H, Xia E-H, Zhang H-B, Gao L-Z. Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf15* function and evolution in angiosperms. *PLoS ONE.* 2013; 8: e59620. <https://doi.org/10.1371/journal.pone.0059620> PMID: 23527231
83. Hoch B, Maier RM, Appel K, Igloi GL, Kössel H. Editing of a chloroplast mRNA by creation of an initiation codon. *Nature.* 1991; 353: 178–180. <https://doi.org/10.1038/353178a0> PMID: 1653905
84. Tsudzuki T, Wakasugi T, Sugiura M. Comparative analysis of RNA editing sites in higher plant chloroplasts. *J Mol Evol.* 2001; 53: 327–332. <https://doi.org/10.1007/s002390010222> PMID: 11675592
85. Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. *RNA.* 2014; 20: 1499–1506. <https://doi.org/10.1261/rna.045575.114> PMID: 25142065
86. Lee J, Kang Y, Shin SC, Park H, Lee H. Combined analysis of the chloroplast genome and transcriptome of the antarctic vascular plant *Deschampsia antarctica* Desv. *PLoS ONE.* 2014; 9: e92501. <https://doi.org/10.1371/journal.pone.0092501> PMID: 24647560
87. Wang W, Zhang W, Wu Y, Maliga P, Messing J. RNA Editing in chloroplasts of *Spirodela polyrrhiza*, an aquatic monocotyledonous species. *PLoS ONE.* 2015; 10: e0140285. <https://doi.org/10.1371/journal.pone.0140285> PMID: 26517707
88. Tillich M, Lehwark P, Morton BR, Maier UG. The evolution of chloroplast RNA editing. *Mol Biol Evol.* 2006; 23: 1912–1921. <https://doi.org/10.1093/molbev/msl054> PMID: 16835291
89. Nagy E, Hegedűs G, Taller J, Kutasy B, Virág E. Illumina sequencing of the chloroplast genome of common ragweed (*Ambrosia artemisiifolia* L.) *Data Brief.* 2017; 15:606–611. <https://doi.org/10.1016/j.dib.2017.10.009> PMID: 29085876